DOCUMENT RESUME

| | | |
|---|---|---|
| ED 428 135 | | UD 032 775 |
| AUTHOR | Mcmillan, James H. | |
| TITLE | Unit of Analysis in Field Experiments: Some Design Considerations for Educational Researchers. | |
| PUB DATE | 1999-02-00 | |
| NOTE | 19p. | |
| PUB TYPE | Reports - Evaluative (142) | |
| EDRS PRICE | MF01/PC01 Plus Postage. | |
| DESCRIPTORS | *Educational Research; *Research Design; Research Methodology; *Validity | |
| IDENTIFIERS | *Unit of Analysis Problems | |


ABSTRACT

Three experimental design issues are examined in relation to the appropriate unit of analysis. Independent replication of the treatments for each subject, independence of observations when gathering dependent variable data, and randomization of groups of subjects are factors that can affect the statistical model and interpretation of results. Examples from two recently published studies are provided, and implications for validity are discussed. (Contains 13 references.) (Author/SLD)

# Unit of Analysis in Field Experiments: Some Design Considerations for Educational Researchers[1]

James H. McMillan
Virginia Commonwealth University
Box 842020
Richmond, VA 23113-2020

February, 1999

[1]Appreciation is expressed to Donelson Forsyth and James Raths for their comments on an earlier draft of this paper.

Abstract

Unit of Analysis in Field Experiments: Some Design

Considerations for Educational Researchers

Three experimental design issues are examined in relation to the appropriate unit of analysis. Independent replication of the treatments for each subject, independence of observations when gathering dependent variable data, and randomization of groups of subjects are factors that can effect the statistical model and interpretation of results. Examples and implications for internal validity are provided.

For more than three decades the seminal work of Campbell and Stanley (1963) has stood the test of time. Clearly, their explication of experimental design and threats to experimental validity has provided the foundation for knowing how to conduct and evaluate applied educational research. While there have been some attempts at revising, expanding, and reconceptualizing the "threats" to experimental validity (Campbell, 1986; Cook & Campbell, 1979; and Krathwohl, 1987, 1993), the original work has been remarkably resilient. What is summarized here is an attempt, albeit a modest one, to suggest an elaboration of Campbell and Stanley that highlights limitations in educational experiments related to unit of analysis.

After reading educational research studies for about 20 years I continue to be perplexed by published experiments that administer a treatment once to each group of subjects or to a few classes or other intact groups, measure the dependent variable at one time to all members of each group together, and then use individual students as the unit of analysis without regard to possible confounding variables or appropriate statistical models. At issue here is whether some fundamental design principles have been violated, and whether such violations affect the validity of the conclusions. It's not as if the issue has been ignored (Bickman, 1985; Edgington, 1985; Glass and Hopkins, 1996; and Hopkins, 1982; Raths, 1967). Indeed, prominent researchers have addressed unit of analysis in experiments over an extended time period (e.g., Cronbach (1976); Glass and Stanley (1970); Lindquist (1940); Page (1975)), though the issue has been more a statistical than design concern. That is, most attention has focused on how the data should be analyzed, while factors that threaten the internal validity of the experiment, such as "the lawnmower effect," as some have called it, have not been stressed (that's when the lawnmower goes past the classroom window while a treatment is being administered and distracts students).

For some reason or reasons, it seems that much of our profession does not think that this is a very important topic for experimental designs. One indicator of this is that few introductory educational research textbooks address issues associated with unit of analysis. Another indicator is that published studies seem to ignore long-standing advice about whether the proper unit of analysis is groups of individuals, such as classrooms or groups within classes, or individual subjects.

In this article I will first summarize the arguments for three design principles related to unit of analysis that are often violated. Then I will illustrate these issues with two recently published studies. Finally, following a list of implications, I'll suggest an elaboration to Campbell and Stanley (1963) that may help ameliorate the problem.

Independence of Treatment Replications

Correctly identifying the appropriate unit of analysis is essentially a methodological issue related to experimental design, even though unit of analysis is directly tied to hypothesis testing. One basic principle is that each replication of the treatment for each subject is independent of the replications of the treatment for other subjects. In a well-controlled psychological study of perception, for example, independence is achieved if each subject, alone and separate from other subjects, is presented with the entire treatment, then the next subject is again presented with the treatment, followed by another treatment for the next subject, and so forth. Each treatment is independent of the other ones. If there are 30 subjects, the treatment has been replicated thirty times to achieve independence.

From the perspective of experimental validity, why is this important? Why not just put 30 subjects in a room and present the treatment? The answer is based on variations that are inevitable in presenting the treatment. That is, each time the treatment is administered, even though it is theoretically the "same," there will be some differences, and these differences may influence the fidelity of the treatment or the responses of the subjects. For example, time of day, enthusiasm in giving directions, confederate fatigue, and a host of other differences can and do occur. When these effects are spread over a large number of replications, they tend to be balanced due to chance and contribute only to error variance. If there is a single administration of the treatment, however, whatever peculiarities exist along with that administration are confounded with the treatment and create systematic error. Such confounding makes it difficult, if not impossible, to conclude that the treatment as planned and not as occurred with other factors accounts for change. This issue is essentially what Cook and Campbell (1979) refer to as construct validity, knowing the nature of the constructs that were responsible for the relationship.

Suppose in a class of 30 students a treatment is presented once to all the students. In this situation, anything confounded with the treatment, like mood of the teacher, time of day, week, or year, interruptions, room, and a host of other extraneous events, will likely influence all the students in the same way, or at least in the same direction. The prospect of something happening with the treatment, then, constitutes a systematic source of bias. This confounding is potentially a major problem with treatments given to classes or groups over a long period of time because the number of potentially confounding variables increases, and with it the chance of committing a Type I error. For example, in much research on cooperative learning the treatment is given to each small group of students, but the idiosyncratic nature of how each group progresses, based on who is in the group, is likely to be a primary determinant of the results. Each group, therefore, should constitute an experimental unit, and be formed prior to assignment to treatment or control.

Thirty students divided into six groups of five, then, should be analyzed as if there were six subjects, not thirty (as long as the six groups are separated from each other). Many field studies give treatments to groups of students and then use individual students as the unit of analysis, without regard for possible confounding factors. This may help explain why educators have had so much difficulty in building a coherent body of professional knowledge. Not only are there opportunities for confounding variables to influence results, which can lead to contradictory findings when treatments are replicated, external validity and power are weak. If there is no statistically significant difference, the lack of power makes it tenuous to conclude that there is, in reality, no difference.

Independence of Observations

The importance of independence is also a concern with the nature of the dependent variable and how the dependent variable is measured. If subjects are tested or observed as a group, rather than individually, there are two reasons for concern. First, individuals may very well influence each other in a group setting when responding to the dependent measure. This could be very explicit, when members of the group openly talk to each other or discuss appropriate responses, or there could be inadvertent signals like verbal outcries (e.g., exclamations or questions) or nonverbal messages (e.g., facial expressions, alertness, body posture, eye contact). Thus, subjects are influencing each other, violating independence and possibly creating systematic error. Second, random differences between the situations or groups that influence how individuals respond are turned into systematic effects. For example, if in one group there is a distraction during the completion of a posttest, such as an announcement or someone coming in the class, this distraction, because it occurs for all subjects, becomes a source of systematic error. If a control group did not experience such an event, then the distraction is confounded with the measurement. Other confounds with the measure of the dependent variable, such as time of day, setting, who administers the dependent variable, and other factors, can also affect the way subjects respond and create systematic error.

The magnitude of the effect of violating the independence of observations principle is contingent on the nature of the dependent variable. In a study which uses standardized test scores as the dependent variable, relatively little variation would be expected due to the high degree of consistency of administration procedures, control over external distractions, and the minimal effect that subjects would have on each other while taking the test in the same room. However, if the dependent variable is student attitudes or self-concept, the effects of distractions, nonverbal cues, time of day, and who administers the instrument, will more easily affect students in a systematic way.

The effect of nonindependence of observations on the unit of analysis is that there are not as many truly independently measured effects as indicated by the number of subjects. Thus, to the extent that subjects are measured as or in groups, rather than individually, the unit of analysis is more accurately ascribed to the group. Essentially the nonindependence creates a "group" effect, rather than allowing the treatment to have independently measured effects on each individual.

Random Assignment

A final factor to consider is how random assignment is carried out. Researchers use random assignment to achieve statistical equivalence between groups and control for many threats to internal validity. This type of randomization creates what is called a true experiment. But the goal of statistical equivalence is reached only when the units randomly assigned, either as individuals or as groups, are the same in number as the units that independently receive the treatment. Flipping a coin to see which of two classes will receive the treatment and which will serve as the control does not constitute random assignment of individuals.because the goal of statistical equivalence has not been achieved. Randomly assigning students to two classes and then carrying out the treatment once in one class also does not constitute a true experiment. This is because, even with random assignment, treatment by group confounds emerge between assignment to condition, administration of the treatment, and administration of the measures. At best these procedures result in a quasi-experimental design. Thus, selection remains as a serious threat to internal validity if two intact classes are assigned randomly, and even if students are randomly assigned to each class, when there is only replication of the treatment in each class, resulting in confounding effects previously discussed.

If the treatment is carried out in each class, then classes must be randomly assigned to claim statistical equivalence. Suppose a study is examining the effectiveness of using small groups to conduct counseling sessions. Each group is randomly assigned to treatment or control, and the treatment is carried out for each group. Clearly in this type of design the appropriate unit of analysis is the number of groups, not the number of individuals.

<u>Examples</u>

Two examples of published research will be used to illustrate how researchers, both primary investigators and journal manuscript reviewers, seemingly ignore the effects of violating principles of experimental design related to random assignment, number of treatment replications, and independence of treatments and observations.

The first example is an experiment in which 89 university undergraduates participated in a study of the value of community service (Markus, Howard, & King, 1993). In this study two of eight discussion sections of a course entitled "Contemporary Political Issues" were "randomly designated as 'community service' sections" (p. 412). One graduate teaching assistant taught the two experimental sections, which contained 37 students; three other teaching assistants taught the other six sections, which contained a total of 52 students. Students had a choice about which type of discussion section (experimental or traditional) they enrolled in. Procedures were employed to control some potential sources of bias, such as differences between discussion section teaching assistants and course evaluations. Dependent variables included self-reports about the importance of volunteering and helping others, attitudes toward people in need, amount learned, and course grades. Pre and post surveys of personal values and orientations, grades, attendance, and end of course self-perceptions about the effect of the course showed statistical significance in the hypothesized direction. Students from the two experimental sections reported more positive effects of the course, greater change in values about volunteering and helping others, better attendance, and higher grades than did students who attended the traditional discussion sections. Students were used as the unit in data analyses.

In this study it is clear that students were not randomly assigned to the six discussion groups, and that designating two of the six sections "randomly" to be treatment groups is an example of using the idea of randomization to infer that the sections were statistically equivalent. To the credit of the authors, some data are presented to indicate this equivalence, but no limitation is mentioned concerning possible bias that could exist because students self-selected into sections (before they knew about the experiment). It is also clear that at least part of the "treatment" is a group experience, idiosyncratic to the doctoral student leading the discussion sections and the time and nature of each of the two treatment sections. There is a potential for much treatment diffusion within the two sections, as students who may be strongly affected by their community service experience interact with others. In essence, then, some of the "treatment" is repeated twice, once for each of the discussion sections. Other aspects of the treatment related to what students experience in the field, could be considered individual replications. Again, however, there is no mention of limitations related to how group treatment effects impact the results. Finally, it is likely that there was not independence of observations. Although not clearly indicated in the article, students probably completed the surveys at the end of the course in groups, susceptible to influences from others or even from discussion section leaders if responses were gathered in that setting.

Despite the fact that other threats to internal validity exist in this study, such as selection, experimenter bias, diffusion of treatment, and statistical conclusion (univariate gain scores were used), the fact that there was little regard for inadequate randomization, the limited number of treatments independently replicated, and nonindependence of observations, suggest further limitations. These issues need to be addressed in interpreting the results and formulating conclusions. At best, we are left with the finding that some aspect or aspects of the entire treatment experienced probably affected the students, but that conclusion is tentative and, because specific change agents can't be identified, we can't be sure what really caused any effects.

The second example is a study of the effect of cooperative learning groups on the achievement and self-concept of economically disadvantaged fourth grade students (Lampe, Rooze, & Tallent-Runnels, 1996). Eight intact social studies classrooms in two schools were used for the study. Two classes in each school were randomly assigned the cooperative learning condition; the other two classes in each school continued a whole-class, textbook-centered, teacher-directed format. Both treatment and comparison class teachers had received training in cooperative learning. Pretests and posttests were administered for both achievement and self-esteem over a twelve week period. ANCOVA, using the total number of students to calculate degrees of freedom, was used to analyze the data, resulting in significantly higher achievement for students in the cooperative condition and no significant differences for self-esteem.

Predictably, the pretest mean for students in the cooperative classes was quite different from the mean of the traditional classes. It turned out that the traditional classes had a mean achievement score of 21.11 (SD=5.02), and the cooperative classes a mean of 24.09 (SD=5.33). Thus, despite "randomization" the students who were hypothesized to achieve more had initially higher scores on the dependent variable. This illustrates what can happen when random assignment is done with intact groups or classes rather than with individual subjects (It can also occur with random assignment of students). Even with covariance analysis, (which many would maintain is not appropriate in this type of study) differences between the intact groups are simply not controlled when the unit of randomization is the class. Other differences between the students in the cooperative and traditional classes related to achievement would constitute rival hypotheses. In addition, there are bound to be other differences between the classes, in such factors as teachers and classroom climate, and it would be very difficult to partial out the effect of these confounds. Because the treatment was "replicated" only four times, any number of specific incidents or events not directly tied to cooperative instruction could well have influenced all or most of the students in a given classroom.

Finally, as in the first example, there is some question about the independence of observations. While testing students individually on social studies knowledge probably meets the independence assumption, since all students in the same class presumably took the tests at the same time, events and influences during the testing may systematically influence performance. Overall, then, this study has some significant weaknesses that are not addressed. Along with some traditional threats, such as diffusion of treatment, selection, experimenter (teacher) bias, and history, there is a strong possibility that factors confounded with the small number of treatment replications and nonindependence of observations could also influence the results. All of these threats need to be considered before we accept the findings as contributions to what we know about the effects of cooperative learning.

Implications

What are some implications of these design considerations for researchers? The first implication is to recognize these factors as contributing to principles of designing experiments and educate researchers about how they impact internal validity. Greater awareness is needed by the research community. There is a need for those of us who are responsible for teaching educational research and advising students to include these factors in our education of future researchers.

Second, we could communicate through our reviews of manuscripts submitted for publication that these factors, when appropriate, need to be addressed. This process would be encouraged if guidelines for reviewers included directions for considering these issues.

Integrative reviews of literature, which depend on a critical examination of the quality of the research, need to use these principles in judging the quality of studies included in a review.

Third, when designing field experiments, the unit of randomization should strictly determine the unit of analysis or statistical model. This means that a study that randomly assigns two of four classes to the treatment has an n of four, not a sample size equal to the number of students in the classes. This designation of n recognizes the quasi-experimental nature of the design and encourages a focus on potential confounding variables. Since it is likely that most experiments will not have many classes or other units to be randomized, a recognition of this principle will alert researchers to possible confounds. Hopkins (1982) and Glass and Hopkins (1996) show how to use individuals as the unit of analysis in such designs by including appropriate random factors in nested statistical analyses. Page (1975) suggests treating each classroom as if it were a single subject, then treat sub-groups of students within the class as if each sub-group represented a repeated measure. In this approach each subject is measured under different conditions. However, the use of correct statistical models, while a definite improvement over using a more simple analysis, does not assure that confounding variables have been accounted for. At the very least, researchers need to address the issue in discussing limitations and plausible alternative hypotheses, particularly the threat of differential selection.

Fourth, researchers need to take steps to minimize problems created by nonindependence of observations. This could include planning so that the conditions under which observations or responses are made foster independence (e.g., room arrangement that would make it difficult for students to see or hear others while responding to the measure), directions that emphasize individual work, and close monitoring when subjects are responding to the dependent variable measure.

Finally, these factors illustrate the criteria needed for a true experiment, one that is considered strong in internal validity. For an investigation to be classified as a true experiment there should be randomization of the unit that receives each replication of the treatment, whether that is the student, group, or classroom, and independence between each administration of the treatment. This implies a definition of "true experiment" that seems to be ignored by many researchers. All too often a study is interpreted as if it were a true experiment when, because of violating one or both of the above conditions, the study is at best quasi-experimental.

Some Suggestions

What can be done to increase the visibility of and application of these principles of experimental design? First and foremost, perhaps, there is a need to change the way "true" experiments are distinguished from "quasi" experiments. It is not sufficient to base the difference solely on the basis of some kind of random assignment. The randomization procedures should be consistent with the desired unit of analysis as well as consistent with the number of independent replications of the treatment. Thus, random assignment of students to three groups, by itself, does not mean that the experiment is "true" if the treatments are assigned to and carried out with groups. There also needs to be independent replications of the treatment with individual students in each of the groups. Independence of replications is needed to rule out extraneous events or other influences that may occur and become confounded with the treatment. This suggests that it is important for researchers to be very clear about how randomization and treatments are carried out.

Second, it may be helpful to elaborate some on Campbell and Stanley (1963). To accommodate the issue of independence in measuring the dependent variable, the definition of instrumentation could be expanded. Instrumentation would include the requirement that each subject or unit responded independently from other subjects or units. Thus, in examining whether instrumentation was a possible threat to internal validity, there is more to analyze than simply whether there are changes in how data are collected from pre to posttest.

To highlight the importance of the number of times the treatment is replicated, it may be useful to consider a new "threat" to internal validity. This threat could be labeled treatment replications to emphasize the importance of understanding whether the treatment was replicated for each subject or for each larger unit of subjects. The question to ask is: "how many times was the treatment repeated?" When the treatment has only been implemented once, or a few times, despite the number of subjects in each group, there is a need to think about whether confounding factors could influence the results, and how the "treatment" should be defined.

Finally, there is a need to use the appropriate statistical model to analyze the results. When classrooms are used to administer a treatment, students and teachers are nested within classrooms, and such a factors can be included in the analysis of variance if the factors can be considered random. Sub-groups can also be used with repeated measures split plot analyses.

Summary

Educational experiments in the field can yield invaluable information about the effect of instructional methods, inservice programs, and other factors on students and teachers. However the applied setting in which the research is conducted creates many design problems and issues. While some of these problems and issues are addressed in discussions of experimental design and internal validity, three seem to be often overlooked - independence of treatment replications, independence of observations, and inadequate randomization. Each of these is an important consideration in determining the appropriate unit of analysis for a field experiment, for identifying possible threats to internal validity, and for using an appropriate statistical model. Hopefully, greater attention to the issues will enhance the quality of educational experiments and increase the contributions of experiments to our knowledge base.

References

Bickman, L. (1985). Randomized field experiments in education: Implementation lessons. In R. F. Boruch and W. Wothke (Eds.) Randomization and field experimentation. New Directions for Program Evaluation, no. 28. San Francisco: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally College Publishing Company.

Cronbach, L. J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Stanford, CA: Stanford University.

Edgington, E. S. (1985). Random assignment and experimental research, Educational Administration Quarterly, 21, 235-246.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology (3rd Ed), Needham Heights, MA: Allyn and Bacon.

Glass, G. V. & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations, American Educational Research Journal, 19, 5-18.

Lampe, J. R., Rooze, G. E., & Tallent-Runnels, M. (1996). Effects of cooperative learning among Hispanic students in elementary social studies. Journal of Educational Research, 89, 187-191.

Lindquist, E. F. (1940). Statistical analysis in educational research. New York: Houghton-Mifflin.

Markus, G. B., Howard, J. P., & King, D. C. (1993). Integrating community service and classroom instruction enhances learning: Results from an experiment. Educational Evaluation and Policy Analysis, 15, 410-419.

Page, E. B. (1975). Statistically recapturing the richness within the classroom. Psychology in the Schools, 12, 339-344.

Raths, J. (1967). The appropriate experimental unit. Educational Leadership, 25, 263-266.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*Educational Resources Information Center (ERIC)*

ERIC

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Unit of Analysis in Experiments: Some Design Considerations For Educational Researchers

Author(s): James H McMillan

Corporate Source: Virginia Commonwealth University

Publication Date: [illegible], 1999

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

**Level 1**

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

**Level 2**

☐ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

*"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."*

Sign here→ please

Signature: J McMillan

Printed Name/Position/Title: James McMillan / Professor

Organization/Address: Box 842020 Richmond, VA 23284

Telephone: 804 828 1332

FAX: 804 225 3554

E-Mail Address: Jmcmilla@saturn.vcu.edu

Date: 7/20/99

(over)